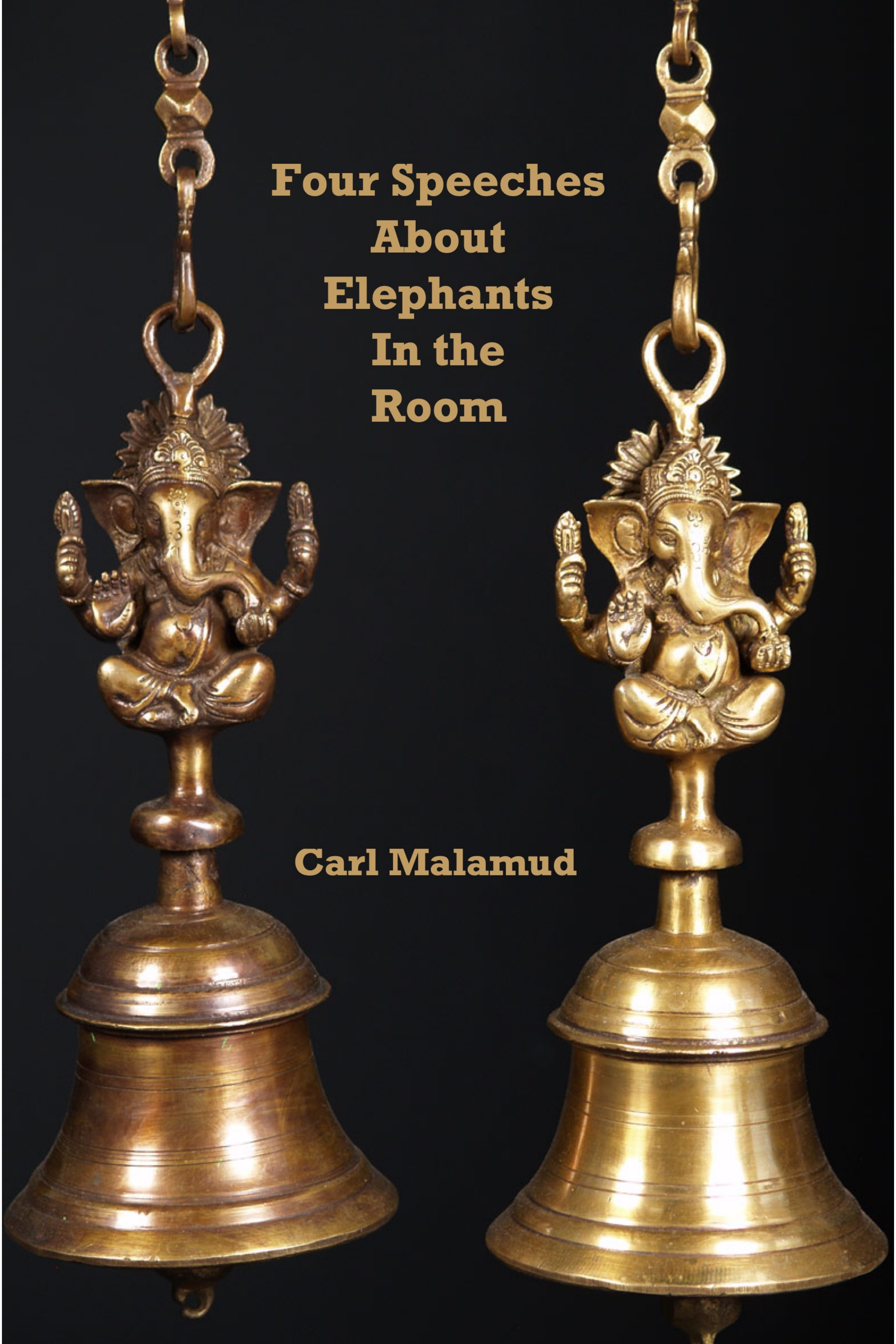

Four Speeches About Elephants In the Room

Carl Malamud

Four Speeches About Elephants In the Room

Carl Malamud

The Rule of Law

The Agami Summit

December 5, 2018

Sheraton Hotel, New Delhi

Good evening. Thank you Sachin for that kind introduction.

I'd like to also thank the core team at Vayam and HumLab for making tonight's Agami Summit possible and for inviting me to speak.

Ladies and Gentlemen.

Our distinguished judges for this evening's Agami Prize. Honorable Chief Justice Gita Mittal.

The finalists for the Agami prize. (A truly amazing group!) My fellow citizens.

We are here tonight to talk about the future of legal innovation in India—to celebrate what is good and great about the Indian legal system, both right now and also what's next—Agami! Tonight is a celebration of "innovations and entrepreneurial initiatives—that can exponentially increase quality, effectiveness, access, and inclusion—in and around law and justice in India."

So, you may rightly be asking yourself why your first speaker is not from India. Indeed, I'm not only a firangi, I have a further confession to make, which is that I'm not even a lawyer.

Although I'm not Indian, I do share a common heritage with you. America and India are the two largest and most vibrant democracies in the world. We are both melting pots.

Both our countries have brought many cultures, religions, points of view, and customs together and out of that we have built democracies that are constitutionally based—that are of the people, by the people, for the people.

Both India and the United States celebrate the rule of law—in both our countries it is the people that are the sovereigns. We—the people—we own our governments.

I just landed here in Delhi a few hours ago from London, where I was meeting with Members of Parliament and civil society groups, attempting to get all legally-mandated public safety standards required in the European Union available for anybody to read—without paying a tax or signing a license agreement first.

It is very hard to get attention in England right now. They have been engaged in what some people in the country seem to feel is a fight for liberation—Brexit—a virulent quest to "Quit Europe."

This has not gone well. It appears that instead of *purna swaraj,* they are going to end up with dominion status and will continue to bow to a supreme law from the Euro Raj in Brussels—but will not enjoy the benefits of true and equal citizenship. Perhaps you can relate?

Personally, I think Brexit was a huge mistake, but then I'm not a British citizen, so this is not my call.

My visit to London to make legally-mandated public safety standards available for all to read—and for all to speak—is also the reason I am here in India.

We will be appearing before the Honorable High Court of Delhi on Friday in our Public Interest Litigation suit to make Indian Standards more broadly available for the students, government employees, engineers, and citizens of India.

As in Europe, in India technical standards—such as building codes —are sold for high prices and restricted by copyright, despite the fact that they have the force of law just like any other government regulation or edict of government.

I'm very pleased to be joined in this PIL by two Indian co-petitioners, including my good friend Dr. Sushant Sinha who is here tonight, and lawyers from the firms of Nishith Desai and Associates and the Chambers of the Honorable Salman Khurshid, who are also here.

The reason we're in court is I committed an act of satyagraha. I bought all 19,000 Indian Standards. These are expensive documents. The National Building Code costs 14,000 rupees. For a book! I bought the standards, and I posted them for free access.

I committed this act, I notified the Bureau of my reasons, they said these laws were their property. I respectfully disagreed and said they belonged to the people. We petitioned the Ministry to reverse their position, and when they refused, we went to court. They said we broke the law.

We said we posted the law. Now it is for the courts to decide.

My efforts to liberate the law—to make it available for all to read and to comment upon, for all to suggest changes—for all of us to know our rights and our obligations as citizens—this is a struggle for all of us, not just lawyers.

Law is the bedrock from which we build the temple of justice. The law is the raw material of our democracies.

Because the U.S. and India are the largest democracies in the world, we have a special obligation to make our system of laws and justice effective and transparent. We face great challenges in both our countries. We also face a bright future, but only if we struggle to make that future happen.

But we must struggle. This is my message tonight. We must not be complacent. Martin Luther King said it so well when he told us that "change does not come rolling in on the wheels of inevitability, it comes only with continuous struggle."

Tonight, we are here at the Agami Summit, the future of legal innovation. Big prizes. Fantastic finalists, dealing with important and transformational issues. Fighting human trafficking. The rights of migrant workers. Legal advice for citizens. Promulgation of the law for lawyers and citizens alike. This is an important event at an important moment.

I am particularly heartened to see here tonight the Honorable Chief Justice Gita Mittal of the Jammu and Kashmir High Court to discuss the vulnerable witness program, as well as representatives from the marvelous Telangana Prison project—both of which are evidence that innovation and change can come from within the legal system.

Change must come from within our legal system, but it must also come from the outside. If we own our government, we must assert that ownership and be the democracy we wish to see.

I have said that the law is any democracy. This is because we are governed by the rule of law. Let me elaborate on that thought.

The rule of law means we govern ourselves by a rule book. We don't make it up as we go along—after the fact, haphazardly, on the whim of some official—we don't make the rules in back rooms filled with smoke. We write the laws down and we promulgate those laws for all to know and we make those laws fair and of general applicability.

These are the three principles of the rule of law. Let me go over them one at a time. I draw here on Lord Thomas Bingham's famous formulation of the three principles in his seminal book titled *The Rule of Law.* It is a short, well-written book which I highly recommend.

The first element of the rule of law is the principle of jurisprudence. This how we make the law, be it opinions of a court, acts of parliament, or the delegated powers of the regulators.

The rule of law says the laws shall be made prospectively—will apply to future actions—and that the laws shall be written down. We do not allow arbitrary decisions to be handed down on the caprice of some official.

As John Adams said, we are a government of laws not of people. An empire of laws, not of people. An empire of laws is concerned with right. An empire of people is concerned with power. Checks and balances, the separation of powers, the mechanics of justice. These are the vital components of the machinery of democracy.

But if all we have is the principle of jurisprudence—of committing the laws to writing and developing them prospectively—we do not yet have the rule of law. For the law to be effective, it must be known. As citizens, we must know our rights and our obligations. Ignorance of the law is no excuse. The second principle of the rule of law is thus the principle of promulgation.

The principle of promulgation of the law goes back to the most ancient times. Throughout India, we see the rock edicts of the great emperor Ashoka, the dearly beloved of the gods. Those edicts communicated the principles of *dharma* to the people.

If you read the rock edicts, you will see that they run from the most general principles of justice—such as religious tolerance—to specific public safety proclamations about the safety of medicines and roads.

In Rome, the Twelve Tables of the Law served the same purpose. When the aristocrats were making up the law as they went along, charging for access to the courts, it was the people, the plebes, who revolted and demanded that the patricians write those laws down and to display those twelve tables in every marketplace.

In England when King John made access to justice a question of access to wealth, the Barons confronted him on the fields of Runnymede and demanded that the King sign Magna Carta.

Magna Carta laid down the general principles of an open and transparent system of justice. It decreed "to no one will we sell, to no one will we deny or delay right or justice."

Like the Ashokan edicts, Magna Carta also dealt at length with technical public safety proclamations having to do with the standardization of weights and measures, to ensure the free navigation on English rivers, and the safety of bridges.

Law is all about the details, which is why I care so much about the details of technical laws such as Indian standards. Indian standards encompass the safety of buildings, the transportation of hazardous materials, the safety of our water and air, the safety of machines in factories toys for children. These are laws that matter to everybody, and they should be made carefully, and they should be promulgated freely.

These two principles I have discussed—the principles of jurisprudence and promulgation—these are not sufficient for the rule of law. You can write your laws down, and you can promulgate them and you do not not yet have justice. All you have is rule by law, not rule of law.

You can make a law that says people of color in the southern states of America must ride on the back of the bus. You can make that law and you can promulgate that law far and wide. You can make it known to all, you can enforce it with dogs and firehoses and police intimidation.

You can say people of color in the south of the United States may not eat at lunch counters with white people. That they may not use the same drinking fountains or the toilets, that they shall go to separate schools, that they shall be denied the right to vote, that people of color may not make their homes next to white people.

You can say all these things, and do so prospectively, and you can promulgate those laws, but you do not have a just society. This is not the rule of law.

In India, the laws said that Indians must grow indigo but must not make salt, that they could not vote—that indenture under cruel and unusual provisions in foreign lands was allowed as a transparent replacement for a system of slavery—the issue that kept a young lawyer named Mohandas Gandhi in South Africa for two decades—where he learned the art of satyagraha and showed us the courage to practice it. This was not a just society. This was not the rule of law.

For true rule of law in a democratic society we must go beyond the principles of jurisprudence and promulgation and embrace the principle of justice. For the rule of law to be true, the laws must be laws of general applicability, not laws for or against one person or one community or one religion. The laws must be fair and just to all.

The principle of justice is the hard part of the rule of law, the hard part of democracy, but we must pay attention to all three principles. We must make the law. We must promulgate the law. We must make good law.

We can clearly see from the efforts of our finalists tonight that the law is there for all of us, not just the legal practitioners. Informed citizens are the key to a working democracy.

An informed citizenry—citizens taking ownership of their government—this is the only way our society can work. This was eloquently stated by John Adams in his *Dissertation on the Canon and Feudal Law,* and I'd like to quote to you a few passages from this classic text.

John Adams said "wherever a general knowledge and sensibility have prevailed among the people, arbitrary government and every kind of oppression have lessened and disappeared in proportion."

He said "liberty cannot be preserved without a general knowledge among the people, who have a right, from the frame of their nature, to knowledge … they have a right … to that most dreaded and envied kind of knowledge, I mean, of the characters and conduct of their rulers. Rulers are no more than attorneys, agents, and trustees for the people."

He said "the preservation of the means of knowledge among the lowest ranks, is of more importance to the public than all the property of all the rich men in the country … Let us dare to read, think, speak, and write."

He said "let the public disputations become researches into the grounds and nature and ends of government, and the means of preserving the good and demolishing the evil. … In a word, let every sluice of knowledge be opened and set a-flowing."

It is easy to be cynical about our governments today. I certainly have grave doubts about the current administration of Donald Trump and the gridlock, ignorance, and corruption that is thriving in Washington, D.C. I know it is also easy to be cynical or despondent about the future of democracy in India.

But there is much celebrate here in India and we must not forget that.

India is constitutionally directed and constitutionally grounded and we must all celebrate the ground-breaking work of the constituent assembly under the leadership of Dr. Ambedkar, there is so much to learn from reading their deliberations.

We must also celebrate the huge body of constitutional law from the Honorable Supreme Court of India and the amazing work of the High Courts. These tribunals have handed down some of the most influential judicial pronouncements in the world.

We must not forget that India has the best Right to Information law in the world, the result of a 25-year struggle by Aruna Roy and MKSS, a model and inspiration for the world, an epic tale of citizens changing how their government works.

Many of you in this room are lawyers, many of you practice the law. But we must also not forget that the practice of law can only have one goal, and that is to serve society.

As I read the history of fight for liberation in India, I am struck by how many of the leaders of this struggle were lawyers who gave up their law practice to become public workers. Sardar Patel. Rajaji. Jawaharlal Nehru. Motilal Nehru. Mahatma Gandhi. And, so many others.

They were lawyers, but first they were citizens. Like them, we must all strive to make our society better and tonight we are celebrating the efforts and dedication of some of our fellow citizens—the finalists for the Agami Prize.

But we must do more than celebrate their achievements. We must work with them, we must struggle with them, we must all strive to make our society one in which "justice flows like the waters so that righteousness pours forth like a mighty stream."

Thank you very much.

Gyan Swaraj: Building A Truly Public Library of India

Shree Guru Gobind Singh Tricentenary University
Gurugram, Haryana, India
July 30, 2019

Thank you Professor Radhika for that kind introduction and to Shri Amogh Dev Rai for the invitation to speak here today.

My friends.

Chancellor Shri Ram Bahadur Rai Ji; Chairman of the Trust Shri Manmohan Singh Chawla Ji; Pro Vice-Chancellor Tuteja Ji; the Director of Academics, Professor Mittal Ji,

Ladies and Gentlemen. Good morning.

It is a great pleasure to be here at SGT University today. I am particularly pleased to be here at a university that was created with a holy cause in mind, a holy cause of propagating the message of Shri Guru Gobind Singh Ji, the great philosopher and social reformer, a message which says "the spread of learning is the best service to mankind."

Learning is indeed the best service to mankind, it is something that you will begin here at the University, but I hope it is but a temporary stop in your path, the beginning of your travel on the road of lifelong learning.

Knowledge is of course essential as you train for your chosen profession and search for economic opportunity. Lifelong learning is essential if you are to teach your children how to go down their own road, for if you do not read, if you do not learn, you cannot expect your children to follow their own path.

But learning is more than the key to earning a good living or being a good example to your family. It is the key to something much bigger, and that is the topic I wish to address today.

Learning is the key to solving pressing problems such as global warming, poverty, and disease. Learning is the key to achieving the goal of making India a $5 trillion economy. Learning is the key to creating a better and more just democracy.

What I wish to talk to you about today is gyan swaraj, why open knowledge—knowledge available to all Indian citizens, knowledge that is freely accessible to all—is the core underpinning of any truly great society.

☸

Before I get to that topic though—about knowledge here in India—I'd like to tell you two stories about my own experiences in the United States. I want to tell you those stories so you can understand my perspective, but also because I believe these goals are universal, that they go beyond India to the rest of the world—and because I would like to explain why I have come to believe that if there will be a revolution in access to knowledge, it will have to start here in India.

For the last 40 years, I have had the privilege—and the honor—to work alongside tens of thousands of my peers on what has become the global Internet, a network of networks that has grown to billions of computers, that connects billions of people. Of course, we must not forget that the Internet still only reaches half the world, and our job must not be considered complete until access to the Internet becomes truly universal.

During the 1980s, when I started working on computer networks, there were many different brands and kinds of computer networks. The idea that any one network might connect the whole world was considered a wild dream, hubris, a dream we all shared perhaps, but one we had to admit was unlikely to become a reality.

The Internet in those days was a network running on a particular set of network protocols. It connected a few hundred computers, then a few thousand computers. There was a file called "hosts.txt" which listed all the computers on the Internet in it. If you wanted to use a modern new service called "e-mail," you'd pull up this file to decide which computer to send your mail to.

But the Internet protocols were not the only game in town. There were commercial networks from companies such as the Digital Equipment Corporation, which they called DECnet. IBM had their own networking solution for connecting big corporate mainframes together, something called the System Network Architecture.

The protocols we use today were developed by a rag-tag group of engineers from around the world called the

Internet Engineering Task Force, the IETF. We met three times a year, but most of our work was conducted by email. By rag-tag group, I mean we didn't even exist: there was no formal corporation or association or society, we just worked together.

There was another set of protocols that was much more formal, and it was called the Open Systems Interconnection model, or OSI. It was developed by the oh so very formal International Standards Organization and it was backed by all the great and the good. IBM supported it, the U.S. Department of Defense supported it, all the telephone companies loved it, the European Commission poured buckets of money into it.

OSI was based on the idea of a "smart network." What that means is that the telephone company would build the network on top of their existing lines and provide users with a number of value-added services.

The only surviving example is the Short Messaging Service or SMS, what you now know of as texting. In the early days of SMS, the telephone company offered the service for 25 cents to send a text, and then the receiver was charged another 25 cents to receive it.

In the OSI model, the smart network would allocate a channel for you. You could pay extra for a high-priority channel. You could pay extra for a secure channel. There was a menu of services, and everything was served a la carte, you paid for everything. There was no all-you-can eat buffet.

The menu had the services that the telephone company decided you could have, you couldn't go back into the kitchen and ask the chef to whip you up something different. You certainly couldn't have your own kitchen, or even a hot plate.

Besides being a so-called smart network (which meant a really complicated network because all those services led to incredibly complicated specifications), OSI had one more attribute. The standards documents that defined the network were published by the International Standards Organization and were amazingly expensive.

I spent thousands of dollars buying these standards to learn how the network worked. The standards were copyright restricted and you couldn't share them with anybody else.

By contrast, the Internet I worked on, the Internet Engineering Task Force protocols, were based on a dumb network. The telephone company was our enemy. Our protocols were based on the end-to-end principle. The only job we gave the telephone network was to take bits in one end and throw them out the other side.

To us the telephone system was a dumb pipe. We didn't even care if the network lost the bits (which it often did), our protocols would keep on sending them until a copy made it to the other side.

The end-to-end principle meant that all the smarts were on the computers on the edge of the network. This meant that anybody could create a new service, like Tim Berners-Lee did when he created the World Wide Web, like the engineers who created our modern email did, or streaming protocols for audio and video, or chat rooms, or file sharing, or whatever bright idea any of you here today might have for a way to change the world.

The other thing that made our network different was that our standards were totally open and our process was open. Anybody could read them for free and anybody could copy them. Anybody could propose a new standard.

We had a strong practical ethos: our motto was "rough consensus and working code." If you wanted your idea to be a standard, you had to convince other folks that what you had in mind worked not just in theory, but in practice.

By contrast, the OSI model was all about theory not practice, and as my colleague Marshal T. Rose often reminded us, the distance between theory and practice is far greater in practice than in theory.

What I learned from the Internet was that open works, and what I learned is that there is always some stranger out there that is smarter than you are. There were many, many occasions in which a standard wasn't working properly and some person nobody had heard of would speak up and say "I know a better way."

Innovation always springs from unexpected wells, you can't micromanage that process from the top. The Internet came into being because it was open. Only open can scale. Only open works.

Let me tell you a second story before I turn my attention to India. In the early 1990s, I was running Internet Talk Radio, the first radio station on the Internet. I ran it as a non-profit organization because I was a big fan of our own public radio and TV networks and because I believe in public work, but also because I wanted the flexibility to do things that weren't radio.

One of the things I did was put the U.S. Securities and Exchange Commission database on the Internet for free access. The system was known as EDGAR, the Electronic Data Gathering and Reporting system. In the U.S., all public corporations are required to file periodic reports.

This is what makes our financial markets open and transparent. Companies have to file quarterly and annual reports. Before a company does an Initial Public Offering, an IPO, they have to file extensive disclosures. Mutual funds have to file reports with the SEC. There are many other such reports that make our markets work properly.

In 1993, those reports were very expensive. If you wanted to read IBM's annual report, because maybe you were a journalist reporting on IBM or a student who might want to work for them, you spent $20 or $30 for each report.

The SEC's model was that ordinary people couldn't consume the basic reports. They weren't smart enough. The reports weren't usable. So, they had set up a system of "value-added." They spent $35 million to put together a scheme where a company called Mead Data Central was the data wholesaler, they "added value" to the raw feed.

They then sold feeds for hundreds of thousands per year to data retailers, who "added value" to the documents and sold them to people. This was a $300 million/year market.

Congressman Edward Markey—who chaired the congressional committee with oversight of the financial industry—asked me one day why EDGAR wasn't available on the Internet for free, and I told him I didn't know, but I'd look into it.

I found no conceivable reason why this public information shouldn't be available to the public. The SEC's argument was that no ordinary person would ever want to read these specialized documents except for a few well-heeled Wall Street fat cats, so why should we subsidize their document needs when everybody could make a nice fat profit selling them EDGAR?

I didn't buy that reasoning. So, I decided to do something about it.

I went to the National Science Foundation and got a grant to buy EDGAR from the SEC and put it on the Internet. Think about this a second. I got a grant from the American government to buy the data from the American government to give it away to the American people.

When my grant was announced in the New York Times, all hell broke loose. The vendors screamed bloody murder. A powerful congressional chairman, Mr. Dingell, threatened investigations as to why the National Science Foundation was competing with the private sector.

Luckily, Vice President Al Gore called the New York Times and called this "a big win for the American people," and that bought me a couple of months, so I scrambled and got that service up and running on the Internet.

My SEC service was wildly popular. We reached millions of new users. Senior citizen investment clubs, journalists, students, academic researchers, day traders, corporate employees all suddenly started reading these reports.

After running the service for a year and a half, I shoved it back down the SEC's throat by saying I would terminate the service in 60 days, that it was the SEC's job to run this system, and 17,000 people wrote to the SEC agreeing with me. When the SEC complained they couldn't possibly get this up and running in 60 days, I gave them our source code, loaned them computers, configured their Internet line, and they were on the net.

Two things happened. First, there was a dramatic change at the SEC. When I was running the service, the computer staff over there were very unhappy. They went around telling people our system was unsafe, it was undermining their value-added retail chain, the Internet would lead to viruses getting introduced, and so on and so on. They were not fans.

But, after they took over the service, they went to their bosses and they got to buy fancy new computers, and they found themselves running the busiest web server in the federal government and they got really happy really quick. It was geek heaven for the IT staff.

The other thing that happened was the retail EDGAR industry totally changed their tune. One of the vendors that was making big money selling SEC documents and had protested the loudest came up to me and said "you know, Carl, I wasn't expecting this, but our revenues went up instead of down!"

What had happened is we vastly expanded the number of people reading these documents, and those that were serious about SEC documents went from my free service to the professional offerings, because they had better search, and a better collection of documents, and integrated into their workflow.

Having the core data be open doesn't conflict with private industry's ability to make a tidy profit reselling government data. They just have to exercise their smarts to do something better with the data instead of creating an artificial monopoly.

The lesson from both these stories is that open works. If you want to build a vibrant economy, you need a core open infrastructure that anybody can use. If you want a city with shopping malls and corporate headquarters and factories, you also need public parks and roads and transportation services.

☸

Let me turn my attention to India now. I'd like to give you a few examples of some of the work I'm doing with my colleagues throughout India.

The first example is technical knowledge. I had travelled to India many times in the 1980s and 1990s, but it was always as a visitor. I finished editing my book about databases on a houseboat in Srinigar. When His Holiness the Dalai Lama wrote the foreword to my book about the Internet World's Fair I helped create, I travelled to Dharamshala to present him with a copy.

What brought me back to India was when I started posting technical laws on the Internet. By technical laws, I mean public safety codes, like building codes, fire codes, plumbing codes. These codes have the force of law, and are essential to the public safety, but all over the world they have been sold for high prices and access and use has been restricted.

My work began in the U.S., posting the building code of California, my home state. These codes cost over $1,000 to purchase, and they are an essential part of California law. Even our municipal building inspectors have to buy the codes. A kid that wants to study for a plumbing or electrician's license has to buy the codes.

I bought those codes, scanned them, and posted them on the Internet for free access. We didn't just buy and scan them though—we retyped many of the codes into HTML so they worked on modern browsers and were accessible to the blind. We redrew the diagrams into the modern SVG format, so you could cut and paste the diagrams into your own documents. We exposed the standards to search engines like Google so you could find them.

This work was not without controversy. In the U.S. and in most of the world, these technical codes are made by private not-for-profit organizations who make model codes, and then advocate for governmental authorities to turn them into law.

The National Fire Protection Association makes the model National Electric Code, then has successfully lobbied all 50 states and the federal government to make that code binding law. Despite that, the National Fire Protection Association claims they have the exclusive right to sell this law, and to decide who may or may not read it, and on what terms.

They were not pleased with my work, and the NFPA has joined with 5 other standards development organizations to sue my organization in U.S. federal court. We initially lost the case on copyright grounds, this was then overturned by a unanimous opinion in our favor from the U.S. Court of Appeals, and the case is still pending.

A related case, over our posting of the official laws of the State of Georgia, has just reached the U.S. Supreme Court. Again, we lost at the District Court level, we won a decided victory in the U.S. Court of Appeals, and now our highest court is going to decide the question of edicts of government.

Who may read and speak the laws in a democracy in order to inform their fellow citizens about their rights and obligations under the law? That is the question the court has agreed to hear, and our position is that in a democracy the law is owned by the people.

The government functions as our trustees, they are there at our bidding, and nobody can restrict our rights to read and speak the laws by which we have chosen to govern ourselves. We must all know the edicts of our government for ignorance of the law is no excuse.

☸

In India, technical laws are created by the Bureau of Indian Standards, a governmental body. We purchased all 19,000 Indian Standards and posted them on the Internet for free access.

Over 700 of them have been transformed to HTML. They are all available in bulk, so you can download all the standards by a particular committee or download everything.

The service is wildly popular with students across India, with local and state officials enforced with ensuring public safety, with architects and builders, with factory workers and consumer organizations, with engineers and building owners, with farmers and headmasters, and with ordinary citizens interested in the safety of the world around them. Millions of people have accessed these standards.

These Indian Standards are no casual publication of the government, they are the best codified technical knowledge about public safety in this country.

This is no incidental process. Standards are rules and regulations created by the government. Two Union ministers, 2 members of Parliament, 5 state ministers form the Indian Standards Council and oversee a process with thousands of engineers, government officials, and professors who volunteer their time to create these standards. The standards are then issued for public comment, then approved as an Indian Standard, and a notice is posted in the Official Gazette.

This is essential information. The National Building Code of India specifies proper exits in case of fire for schools, hotels, homes, and other structures, yet it is sold for 14,000 rupees per copy and comes with strict admonishments that you may not copy the building code or even small parts of it.

Indian Standards go far beyond the National Building Code. Many products may not be sold in India without a certification stamp from the Bureau of Indian Standards. In fact, the vast majority of the BIS revenue is from certification, sales of standards is a small fraction of that.

There are many other standards. Frequently in India, we read about workers cleaning sewers dying. But Indian Standard 11972, the "Code Of Practice For Safety Precautions To Be Taken When Entering A Sewerage System," explains what every single worker entering these dangerous locations must know. Yet, the document is not widely available.

There are fairly frequent explosions and accidents in chemical laboratories, but it is exceedingly difficult to find

Indian Standard 4209, "Code of Safety in Chemical Laboratories."

We know of the great damage caused by typhoons, but have you read Indian Standard 15948, "Guidelines for Improving the Cyclonic Resistance of Low Rise Houses and Other Buildings/Structures"? If you live in Odisha, this would be quite relevant!

After I posted those standards, I sent a letter to the Bureau informing them of what I had done, and offered to give them the HTML we had created and work with them to make standards more accessible.

I suspect I could have simply run the service and they wouldn't have even noticed, but it is an important principle of satyagraha—and have no doubt on this point, posting Indian Standards is an act of satyagraha—it is an important principle that one does not sneak around, one must be forthright.

Before Gandhiji marched to the sea to make salt, he sent a letter to the Viceroy, which famously began "Dear Friend." He told the Viceroy of his actions, and invited him to do the needful, but he did not. It was only then that Gandhiji marched to the sea.

Gandhiji was quoting Justice Ranade when he told us that we petition for two reasons. First it is to warn our rulers, but it is also to inform ourselves of our condition. We must educate ourselves before we may act.

After the Bureau objected to my actions, we petitioned the Hon'ble Ministry of Consumer Affairs, Food and Public Distribution and submitted affidavits from distinguished Indian professors of Engineering. We submitted lists of all the government regulations that used the standards. We documented before and after examples of how we transformed the standards to make them more readable for all consumers, and especially more accessible to the visually impaired.

The Ministry rejected our petition, so we filed a Public Interest Litigation writ before the Hon'ble High Court of Delhi. I am joined by two Indian co-petitioners, Sri Srinivas Kodali of Hyderabad, who is a noted activist on topics such as Aadhaar, and Dr. Sushant Sinha, the creator of Indian Kanoon, the prize-winning site that provides all court opinions for free to the public.

We are represented before the court by Sri Jawahar Raja, who successfully argued the Delhi University copyshop case and Sri Salman Khurshid, the former Minister of Law and Justice. The case is pending, and we appeared yesterday before the Hon'ble Court.

As I said earlier, selling the standards is not a hugely profitable or large business for the Bureau. The big money is in certification. But it is even more than that, it is about more than the money.

When people die in horrific fires, or suffer poisoning from improper application of pesticides, the costs to society in money and anguish is immeasurable. The purpose of public safety codes is to serve a larger purpose, to make our society safe, to make our economy function properly, to educate our engineers and farmers and construction workers, to enable our governments to regulate with knowledge.

Selling a standard is being paisa wise and rupee foolish. They must serve a higher purpose.

☸

Let me turn to a second example of work in India, the posting of books on the Internet for all to read.

There was something called the Digital Library of India, a project that was done under the auspices of the Government of India. They scanned 5.5 lakh books from libraries in 10 different scan centers. These are books in 50 languages and are a very unique collection of materials.

A few years ago, I noticed this collection. It was on a pretty awful server. It kept going down. It was really hard to use. Nobody answered the email. So I started making a copy of the files. I managed to make a copy of about 450,000 books.

I put those books on the Internet Archive. The Internet Archive is a non-profit organization, just like my own NGO, Public Resource. The Internet Archive is strictly non-commercial, they have over 5 million books on-line for free access, and another 10 million documents like newspapers.

There are lots of other resources there, including 5 million video files and 7 million audio files. They also run the Wayback machine, an archive of the entire World Wide Web since 1996.

I use the Internet Archive as my cloud for a number of reasons. It's strictly non-commercial, they never charge for content. It's an open architecture, so not only is there a great search engine and a user interface for end users, for developers like myself there are command-line tools that allow me to manage very large collections. I have several crore objects on the Internet Archive.

Most important though is that I can put data up for bulk access by others. If I upload 8,000 books in Tamil, another user can issue a simple one-line command and download all 8,000 of those books. It might take a while, because that's a lot of data, but bulk access is part of the design, not an afterthought. So, I view the Internet Archive as an initial loading dock.

Well, I uploaded these books from the Digital Library of India to the Internet Archive, and then the government server went down. It no longer exists. So, ironically enough, I have the only copy of the Digital Library of India. I have offered to make disk drives of this collection and bring them to India with all the transformed data and give a copy back to the government.

Now, this collection is, to be quite frank, not so good. The scanning was not done well. Pages are missing. The metadata is bad, titles are wrong. There are books in 50 languages, but the titles were all entered in Roman script, not native scripts.

Most importantly, the government was very sloppy on copyright. There were lots of books there they shouldn't have scanned. I was able to get 4.5 lakh books off their servers. I pulled the metadata up into spreadsheets and knocked out 50,000 books that were clearly in copyright. Things like Oxford University Press from the 1970s. I don't know what they were thinking when they scanned those books and simply put them on line.

That left me with 4 lakh books in my collection. There were still a few books in copyright, but that's always going to be the case with a large public digital library. The key to handling those issues is to respond to incoming queries. If somebody writes to the Internet Archive and says "hey, you have my book!" they get a quick answer and if the book is in fact in copyright, it is immediately removed from public view.

For the collection of books I maintain, we get very few of these copyright takedowns, and we've had over 5 crore views on the collection and it is well exposed on Google and other search engines.

We've supplemented that initial collection by harvesting other web collections on the net. For example, we mirrored 23,000 books from the West Bengal Public Library, and 8,000 books from the Tamil Virtual Academy.

For some of the languages, such as Telugu and Kannada, volunteers have been going through the collection and retyping all the titles and creators into native scripts.

There are other resources we've harvested from the net. With my friend Dr. Sushant Sinha, we've created a collection of 4.5 lakh files of Official Gazettes of India. This is the Union government, but also 17 states. And, unlike the sites run by the Government of India or the states, our archive is searchable. You can easily search on metadata, such as show me all the gazettes from Rajasthan or Kerala from the last week.

For languages—like English—where the Internet Archive does optical character recognition, you can also search inside the texts. For example, if you search the collection for the phrase "Vice Chancellor" you will find all 7,144 issues of Gazettes from 2009 through 2019 that have announcements pertaining to Vice Chancellors.

Dr. Sinha has developed a new program which we're in final testing on that allows us to start adding optical character recognition for most Indian languages. It's a pretty cool program, we pull the files out of the Internet Archive, bounce them off the OCR on Google Vision, then shove them back into the Internet Archive and recreate all the files they would have had if they had done the OCR themselves. It's a hack, but it is working great.

☸

The Digital Library of India, Official Gazettes, West Bengal Public Library, and much more are all harvested from the Internet. But, as I said, many of the scans are pretty awful. And, there are lots of materials that are not available on-line yet that really should be there.

There's a group of us that work in technology who are passionate about making materials available online that have banded together in an informal group we call the Servants of Knowledge. We have a high-end scanner installed at the Indian Academy of Sciences in Bengaluru, which has been highly supportive of our efforts.

Using that facility, as well as scanning facilities in the U.S., we've put a few thousand books on-line. The Bengaluru facility has done over 2 lakh pages. There are lots of books in Kannada and Malayalam, books in Sanskrit.

There is a collection I curate called Hind Swaraj that has the complete works of Gandhiji, Pandit Nehru, Dr. Ambedkar, and lots of other things, like the correspondence of Sardar Patel, books by Radhakrishnan, Rajaji, Tilak, Netaji, Gurudev, a huge number of other works by and about Gandhi, and much more.

The scanner we use there is called a Table Top Scribe, it is the same device that the Internet Archive uses to scan 1 million books per year. It has two high-end cameras. You put the book in a V-shaped holder, and you press down a pedal. That raises the book up to the glass, the cameras take pictures of each of the pages, you let the pedal down, flip the page, press the pedal again.

Once you get good at these Table Top Scribe scanners, you can do 800 pages an hour of very high quality. This operation of the pedal strikes me very much like a spinning wheel, and our motto for the Servants of Knowledge is "scanning is the new spinning."

You may be familiar with Gandhiji's passion for bread labor, the idea that you should do manual labor every day. He took inspiration from Tolstoy, and they both took inspiration from the Sermon on the Mount which proclaimed that we should all "earn our bread in the sweat of our brow."

When you think of bread labor, you may think of the spinning wheel, and those iconic pictures of the Mahatma at his charkha, spinning thread, making khadi. But what you may not know was that bread labor was originally something else.

When Gandhiji decided to form his first ashram at Phoenix, the first thing they did as they moved out of Durban was to dismantle the printing press. They put the press on 4 wagons, each pulled by a team of 16 oxen, and they hauled it out to the Phoenix Ashram.

There was nothing there at the time, just wilderness, no houses. But before they built shelter for themselves, they built a home for their printing press. At the Phoenix Ashram, everybody did typesetting, every day. Typesetting was their bread labor.

They used that printing press to pull in articles from all over the world and published them in newsletters like Indian Opinion. Today, you'd call Indian Opinion a blog. Gandhi was a prolific writer and publisher, satyagraha was his most famous tool, but it was the diffusion of knowledge to educate his peers that also an essential a tool.

We now have three of those high-end scanners in India, and my deal with the Internet Archive is if we can make all three of those sing, scanning lots of stuff, they'll us give us three more. The idea is a decentralized operation. We hope others will learn from our model and begin scanning whatever materials that are their passion, their history, their culture.

Scanning is the new spinning.

I have one last example to give you of work in India, after which I will take a step up and conclude with some thoughts as to why this all matters. I've described how we take print materials (and I should mention we also take movies and audio) and scan and harvest them and put them on the Internet Archive.

That's a railroad train of content, a whole set of steps starting with scans, but including metadata and OCR and search engine optimization and then finally hosting with a search engine and user interface so you can access that content. But, that's not the end of the journey.

The content needs to flow back into India and be made available on many different platforms and in many different styles of use. When I come to India, I bring a bunch of little 2-terabyte drives that have a gold label on the front with a picture of Gandhi.

I call those Gandhi drives and they have copies of the Hind Swaraj collection, government books I've scanned, and much more. I give them to vice chancellors and headmasters of schools when I give speeches, and ask them to put the data on their local servers. I'd love for the government to take one of these drives and put their own content on their own servers.

That's an ad hoc effort, but over the last year, we've built something more substantial, working with my colleagues at Jawaharlal Nehru University and IIT Delhi. We started at JNU and built what we call the JNU Data Depot. It has 576 terabytes of disk on two large servers. Another 250 terabytes of disk is spinning at IIT Delhi.

There are two things we want to do with the data depots. One goal is to take the public resources we have on the Internet Archive and put them in India on these depots, and connect them to the National Knowledge Network, so any university in India can quickly and easily download data in bulk. We're just getting started on that capability, but it has great promise.

There is something else we are doing on the JNU Data Depot which is quite significant. There was an article just last week in Nature, the leading science journal in the world, about how India is leading the world in providing a significant new research facility based on our efforts.

One of the servers has been separated from the Internet. It is what we call air-gapped. It is carefully secured in the machine room, and we've locked it down with very limited access. On that server are 73 million journal articles.

Now, many of these articles are in copyright. We can't give you copies of those articles. There are pirate sites like Sci-Hub out there that let people get copyrighted articles, and those sites are very controversial. But, journals cost so much money, most universities can't possibly afford to subscribe to the scientific information that their students and researchers need to pursue their education and the progress of science.

In India, the top universities spend over 140 crore on access to journals and I've heard that when you add in all the universities and labs, India spends 1,400 crore on access to journals.

That's a huge amount of money and despite that, access is incomplete, so people in India and all the rest of the world use sources like Sci-Hub and Unpaywall and Research Gate. India is the second largest user in the world of Sci-Hub, but the U.S. and the U.K. are in the top ten because even at rich schools like Harvard, you can't get access to the literature of science you need to pursue your education.

We're not trying to solve the public access problem. At the JNU Data Depot we're doing something different. We're allowing scientists to do text and data mining on the collection of articles.

This is called non-consumptive use. They are not consuming the articles by reading them and they are certainly not distributing the articles. They are mining them for facts. There is no copyright on facts. Non-consumptive text and data mining does not violate copyright.

For example, we have a researcher that is looking at the secret language of plants. Plants communicate with each other and with other species using chemicals. She had previously mined a small collection of openly available articles on PubMed and built a database of the name of the plant, the chemical, the effect, and the geographic location. It has been very popular with those that research botany, and using our database she hopes to get significantly better results.

Another example is some research coming out of MIT. They are trying to predict which areas of science are going to be important in the future, a tool that would be invaluable to guide policy makers funding research, corporate researchers and venture capitalists, and even young scientists trying to understand where to aim their careers. Again, by using full text on all articles instead of just searching abstracts or metadata, they will get much better results.

☸

The JNU Data Depot is a bit controversial. You may be asking yourself, but what about copyright, how come you didn't ask permission?

Copyright is a bundle of limited rights for a limited period of time. It is not an absolute right. There are many things you can do with articles, even if they are in copyright.

If you are blind, I can give you any copyrighted materials without exception. There is an international treaty that has codified that right across the globe. Copyright is not absolute.

In India, if you are a student and I give you an article "in the course of your instruction," that is an exception to copyright as was so eloquently explained by Justice Endlaw of the Hon'ble High Court of Delhi in the famous Delhi University copyshop case. Copyright is not absolute.

It is well accepted that a library can buy a book and lend out that book. At the Internet Archive, they do controlled digital lending: they purchase a copy of a book that is under copyright and scan it. The physical book is locked away in the warehouse.

They then lend out the digital copy to one user at a time, using digital rights management software so you can't make additional copies. When the user is done reading the book, they check it back in, and the Internet Archive loans it to the next reader.

There is also a concept called fair dealing and there are exceptions to copyright when the materials have been transformed, used as raw material to create new knowledge. In the U.S., Europe, Japan, and other countries there are specific text and data mining exceptions to copyright.

This is because if you prohibit text and data mining, you are holding back the progress of science, and policy makers all over the world have agreed that would be against the interests of society. Copyright is not absolute.

With text and data mining, you are transforming the original text into other things. For example, you might create a digital signature of every word in every article and use that to detect plagiarism. You are not publishing the words, you are publishing a one-way mathematical transformation of the words. Again, not a copyright violation.

Copyright is a limited sets of rights. There can be no ban on the sharing of knowledge. The Nitisatakam teaches us that "knowledge is such a treasure which cannot be stolen." We must take that teaching to heart.

☸

It is my contention that private publishers, particular in the field of science, have over-reached. They would put it to you that they have absolute rights, and they maintain it is up to them to tell you what you may and may not do with ideas.

This is wrong, this is the colonisation of knowledge. A company such as Reed Elsevier acts as a self-styled East India Company of Science when they tell scientists they may not stand on the shoulders of giants without paying an exorbitant toll, that they not promote the increase and diffusion of knowledge through text and data mining.

Because knowledge has been colonised in today's modern world, scientists have become the new Indigo farmers. They take out grants to grow their raw materials and they ship them off overseas, ironically enough in many cases to the United Kingdom. Then, they have to buy back high-priced finished goods.

What is Sci-Hub but an unlicensed salt factory on the edge of the ocean of knowledge? Perhaps Sci-Hub is illegal under current laws, but it is clear that Sci-Hub is of great use to the people of India and the students of the world. Salt and knowledge are both essential to life.

One cannot tell people they may not have knowledge, one cannot tell people they may not have salt. If that is what our world looks like, then our world must change.

☸

Now you may say that universal access to human knowledge is perhaps of importance to a few college students, but what about the real problems facing our world?

What about the crisis of global warming and our neglect of that crisis? Rising temperatures and typhoons. What about the crisis of pollution? India has half of the fifty most polluted cities in the world. Respiratory illnesses are fast overtaking malnutrition as a leading cause of death.

What about poverty? India has a surplus of food yet 200 mllion people are dying of starvation. What about the crisis of disease, those $1000 pills that cost pennies to make and could cure fatal illnesses, if only we could afford them?

And, what about the growing economic disparity in India and in the world? All over the world, the rich are getting richer.

I put it to you that access to knowledge is the first step. It is the first step towards a $5 trillion economy, because that goal depends on innovation and education and that means we must empower all youth and all bright minds, not just those born into rich families. You can't predict innovation, it always comes from the must unlikely places.

Access to knowledge is also crucial to building the political will to solve the problems our governments neglect. Democracy is based on an informed citizenry, it fails if we do not inform ourselves and each other. If we are not tacking global warming in the capitals of the world, that will only change if the people of the world stand up and say enough, we must do better, we must fix this problem.

☸

Do you know what made America a scientific, educational, and cultural powerhouse at the end of the 19th century? We had many things wrong in our country in the 1800s. Slavery. Racism. Corruption in government.

But, we did two things right. First, there was a commitment towards universal education. That commitment was sometimes not carried out in practice, witness the dismal state of the schools for people of color. But the U.S. created the country with the highest rate of literacy, the largest reading public in the world.

The second thing was a flood of cheap books all over the country, and a policy of promotion of the dissemination of knowledge through cheap postal rates that encouraged the proliferation of newspapers.

We did many things wrong, but one thing we did right was the dissemination of knowledge. It made America what it is today.

You can see the same thing here in India. The Bengali renaissance helped lay the groundwork for the swadeshi movement, for the call for swaraj. Gandhi was a prolific publisher, read Young India or Harijan and you will be amazed at how much he wrote.

In Tamil Nadu, the passion for the Tamil language stoked the fires of liberation, led by scholars like Rajaji, in Mumbai look at the profound learning of Lokmanya Tilak and the role he played in disseminating information and ideas.

The liberation of India began with the propagation of knowledge.

Changes comes from knowledge. That change can be revolution, or it can be evolution, but when knowledge gets colonised and limited and rationed, you don't get evolution you get stagnation. John F. Kennedy once said "those who make peaceful revolution impossible will make violent revolution inevitable."

I put it to you that if there is to be a revolution in access to knowledge, a peaceful revolution, it must come from India. Gyan swaraj means we must reject the colonisation of knowledge, just like India rejected colonisation and rule by the raj and showed the way to freedom for the rest of the world.

There can be no $5 trillion economy if India must beg permission to access knowledge in order to educate her youth, to promote new businesses, to solve pressing problems in society, to promote science and invention and culture and history and language.

A first step towards gyan swaraj is the creation of a true public library of India. There are vast treasures buried in the libraries of India, there are vast treasures inside the Government of India that lay fallow and unused.

It would not be unreasonable to set a national goal of creating a truly public library of India, to digitize 3 million books a year for a decade and create an open access repository of 30 million books and other resources. The sums expended would be minor in the national scope of things and the benefits would be immeasurable.

Such an effort would have to be decentralized, one cannot depend on one government agency to create something so mammoth, it must be created by the people. But, the government can do its part. There is so much government information that could be made available much more broadly to help kick-start the effort.

Why are technical standards locked up? Why are the archives of All India Radio not available to all? What about the vast and amazing cultural resources in the Indira Gandhi National Centre for the Arts? The scientific resources of CSIR and the Indian Council for Agricultural Research and other research powerhouses? The comprehensive language resources in the Central Institute for Indian Languages?

☸

I want to close on a personal note, about you and about me. When you hear the words gyan swaraj, you may think this is a national goal, something they do in Delhi, something that politicians make happen.

This is a vending machine style of government. You deposit your taxes in the slot and the government delivers you a product.

But, you own your government. You are the raj in a democracy. Gyan swaraj must start with you.

When you hear the phrase "be the change you wish to see," you may think that is advice one gives to others, about how they should apply themselves. But Gandhi was always very clear that this was advice one has to give to oneself. Swaraj is about self-rule for yourself. Gyan swaraj is about your own personal commitment to learning and to knowledge.

Swaraj is also about something else. It is about public work. The liberation of India came about because many successful people gave it all up for a broader cause. Successful lawyers like Rajaji and Sardar Patel and Motilal Nehru and Gandhiji quit their jobs and devoted themselves to public work, to uplifting the people of India.

Why am I working in India on the cause of access to knowledge? It is because there is a tradition of public work here. I look at people like Aruna Roy, who gave up her plush job in the IAS and moved to a village in Rajasthan and spent the next 25 years laboring for the people, culminating in the Right to Information Act, the most powerful such law in the world when it was enacted.

India has a crying need for knowledge, but so does the whole world. I believe India has a tradition of public work, of tackling difficult problems, that makes this peaceful revolution possible here.

But, this will only happen if we embrace goals of gyan swaraj, if we take up the tradition of public work that created the modern India, and lead the way not only for India, but for the world.

Thank you very much.

Jai Gyan. Jai Hind.

Scanning is the New Spinning

Internet Archive Anniversary Celebration

October 23, 2019

San Francisco, United States

Good evening!

We still do not know who burned the Library of Alexandria.

Perhaps it was Julius Cesar—who was trysting with Cleopatra but feared an attack from her brother Ptolemy. Cesar set fire to the wharves to prevent enemy boats from landing—and that fire spread to warehouses that kept books—and maybe it also burned the library.

Even if Cesar did not destroy the library, other invaders came and went—there were more fires—and we know a few hundred years later this magnificent temple of knowledge was gone.

In India, the great libraries of Taxila and Nalanda had been destroyed by then. The world was entering a bad time. India turned onto itself. Europe descended into poverty, intolerance, and ignorance.

It was the dark ages.

But in Arabia, the great Abbasid Caliphs built the House of Wisdom—the Grand Library of Baghdad—which kept knowledge alive. The texts of Greece and Rome and India were translated into Arabic, they were copied—and they were expanded.

The Hindu numbering system was adopted by the Arabs, indeed you probably know them as "Arabic" numbers—and that was then expanded into new knowledge—a breakthrough—al-Khwarizmi's seminal Al-

Jabr, a book whose title means “a reunion of broken parts”—what we now know of as Algebra.

The great translation movement in the House of Wisdom preserved the knowledge of the ages. Those texts went back to India and they also went to Spain, where the Moors had invaded Andalusa, which was the coal that was stoked into a fire.

Western scholars moved to Andalusa and learned Arabic just so they could translate the works of Aristotle and Archimedes into Latin, but also the great medical, astronomical, and mathematical treatises of the Muslim scholars. This was how we got our renaissance in the West.

The renaissance only came to be because the West stood on the shoulders of the Arabs, who in turn stood on the shoulders of the ancients. Those giants don’t walk the earth, but they live in our libraries.

It was by the replication of knowledge that knowledge was preserved.

☸

Today, we have a new House of Wisdom on the Internet. The Internet Archive is the largest public library on the net, a vast trove with millions of books, video, audio, a complete archive of the Web, and so much more.

The Internet Archive is more than a temple of knowledge, it is a platform that we may all stand upon. It is a library that belongs as much to the readers and learners and developers of the world as it does to the amazing staff that keeps the Archive running.

This is a new kind of library, one that can unleash the great untapped potential of the Internet. Universal access to all human knowledge is the great promise of our times, and the Internet Archive is making that promise become real.

☸

In India, we are using the Internet Archive to house copies of the Official Gazettes, the newspapers of government. We are using external sources to add optical character recognition in Indian languages to make them searchable.

Our volunteers—who go by the name "The Servants of Knowledge"—are crowdsourcing better metadata on books, we are scanning scads of important materials in our facility at the Indian Academy of Sciences and with our partners throughout India.

We have posted 19,000 Indian Standards which are being used by millions of students. We are making the entire scientific corpus available for non-consumptive, non-commercial, text and data mining.

☸

The Sermon on the Mount teaches us to "earn thy bread by the sweat of thy brow." The Bhagavad Gita says those who do not sacrifice for the greater good "eat stolen bread."

Bread labor was central to Mahatma Gandhi's philosophy. Bread labor means you should do something every day by the sweat of thy brow.

When you think of bread labor and Gandhi Ji, a picture no doubt springs to mind of Bapu and his spinning wheel, making kadhi. By empowering the people to make their own cloth, the Mahatma was battling the colonization of his country by the Raj.

What you may not know was that before Gandhi Ji started to spin, bread labor for him was printing. At the Phoenix Ashram in South Africa, everybody had to typeset every day. The dissemination of knowledge was their bread labor. He spread knowledge to fight colonization.

☸

Today, it is knowledge itself that has been colonized.

Scholarly knowledge corporations function as latter-day East India Companies. Scientists are the new Indigo farmers, shipping their raw materials off to journal factories in England, then forced to purchase high-priced finished goods in the form of locked-down journal subscriptions.

What is Sci-Hub but an unlicensed salt factory on the edge of the ocean of knowledge, operating in defiance of this knowledge raj?

There are many factories on the edge of this ocean of knowledge, it is one of the miracles of the Internet. All too many of them though are corporate concerns.

But there is also a public factory, one operating at scale. The Internet Archive is also built on these shores.

The Internet Archive scans millions of books in its huge scan centers in Hong Kong and the Philippines. Volunteers throughout the world scan and upload. It is a tidal wave of knowledge.

Scanning is the new spinning.

The increase and diffusion of knowledge is bread labor that Gandhi would instantly recognize as public work—knowledge is not some corporate asset—we cannot own knowledge, we can only be trustees for the benefit of all humanity.

If we wish to combat the colonization of knowledge, if we wish to create a House of Wisdom—if we wish to realize the great promise of our times of universal access to all human knowledge—this will not happen without you.

This will only work if we all do bread labor.

☸

We must of course thank Brewster Kahle and the staff of this gorgeous edifice they have built—a platform for the ages—a gift they have made for us all. But this is a gift that we must pay forward.

You can scan, you can code, you can upload, you can contribute in any way you see fit—but you must act—for the future of knowledge is up to you—it is up to all of us.

If we want to see an Internet Archive that lasts for eternity, we must be the library we wish to see.

Help save the world. Be a librarian. Help make universal access to all knowledge a human right.

Thank you very much.

The Elephant in the Room

National Frontiers of Science

Indian Young Academy of Sciences (INYAS)

November 6, 2019

Jaipur, India

Thank you Dr. Yadav and the Indian Young National Academy of Science for your kind invitation to be with you.

Honored guests. My friends. Good evening.

There is an elephant in the room with us, and I believe we cannot continue to pretend this is not so. We must confront the issues it has raised, and that is the topic I wish to address.

Knowledge has become colonized. This trend is true for knowledge in all fields, but it is particularly true for science. Your work has all-too-often become the private property of a few large corporate concerns, immensely rich for-profit trading companies that have claimed exclusive control over the corpus of scientific literature.

In these modern times when the Internet makes universal access to human knowledge possible, it is ironic that the scholarly literature has become less and less accessible.

Prices for journals have soared while the costs of production have plummeted. Physical copies have been replaced with digital files that are locked down with digital rights management, terms of use, obscure user interfaces, and aggressive and unfriendly gatekeepers who continually remind scientists like you to keep off the grass.

In this world we have come to inhabit, scientists have become the new Indigo farmers. You ship your preprints and other raw materials off to London where modern-day

East India Companies like Reid Elsevier force you to buy back high-priced finish goods.

These merchant trading companies of knowledge—both for-profit and sadly even many purportedly not-for-profit scholarly societies—these companies are the elephant in the room. They have become a knowledge Raj.

As an author, you are told you may not share your own work without permission, and such permission often comes with a tax and severe limitations. You are told you may not make any use of the scholarly corpus without first applying for a license, and such licenses are often arbitrarily denied.

You may not even make copies of the so-called "version of record" of your own articles for your family or your students, at least according to the fear, uncertainty, and doubt the gatekeepers spread among you.

I have two problems with the idea that an East India Company of knowledge can tell you what—and how—you may read in the course of your pursuit of scientific knowledge.

The big problem of course is that the students and scholars of the world do not have access to the scholarly corpus in order to further their studies. I am pretty sure if I asked you how many of your students used Sci-Hub to do their research, almost every hand would go up.

That's a big problem. A huge problem. A moral travesty.

Education is a fundamental right in our society, the way that people from any walk of life who have the capability to learn may earn a better living, teach themselves a craft or an art, become a professional, or—like all of you here—to practice science to further the increase and diffusion of knowledge.

But, that is not the problem I wish to speak to you about. We are gathered here to discuss the frontiers of science. One of those frontiers is what you may know of as "big data" or "machine learning" or "text and data mining." We use computers to examine the work of those that came before us so that we may stand on the shoulders of giants and reach for new heights.

As scientists, text and data mining offers some amazing opportunities. To conduct this research, however, you must have access to the scholarly corpus. Sci-Hub has approximately 75 million journal articles, Crossref lists approximately 100 million objects with Digital Object Identifiers. Your research labs and universities have access to only a fraction of this body of knowledge.

This second problem—the use of text and data mining on the scholarly corpus for the purpose of scientific research—is a pressing issue for our times. If a scientist believes she can perhaps better cure cancer if she is able to search the text of previous research, then it would be immoral for private parties to tell that scientist she may not proceed upon this inquiry.

But, that is what happens today. Text and data mining is purportedly prohibited. By prohibited, I mean that there are people who make a lot of noise about why this can't possibly be allowed without unjustified monetization and licenses—without clearing the details of your inquiry with them before you may proceed.

Let me give you the example of Max. Dr. Maximilian Haeussler is a researcher at the University of Santa Cruz Genomics Institute. He is using text and data mining to search for references to chromosomal locations in scientific articles, then makes those available in a genome browser. This genocoding software is 200 lines of Python code that searches texts for different ways to refer to a chromosomal location, such as gene symbols, SNP mutation identifiers, or cytogenetic band names.

Max put together a letter requesting permission to crawl all articles on a publisher's site that were published after 1980 (which is the advent of routine reporting of DNA sequences). The code only pulls out 200 character snippets around the match, it is clearly non-consumptive, by which I mean people are not reading or disseminating the article, they are using computers to extract a very small portion.

He sent the letter to 43 publishers. All 43 specifically prohibit crawling their site in the terms of use. For 28, he got some form of partial permission, but in many cases that permission was empty—no site license was forthcoming and technical measures have prevented crawling the site. Fifteen of the answers were an outright no or they simply ignored him.

He has been unable to complete this important work. He has been blocked because gatekeepers don't approve of his research.

A second example of text and data mining is your own Professor Gitanjali Yadav. She is doing a fascinating research project that is examining the silent language of plants. Plants communicate with each other and with other species using chemicals. Each plant has a chemical fingerprint, a unique bouquet of scents and emissions.

The text and data mining she is conducting consists of searching journal articles looking for the names of plant species and their parts, then extracting the names of any volatile compounds associated with those plants as well as details such as where they were reported and the date.

This work began 10 years ago when she searched open sources such as PubMed and created the Essential Oil Database. The database presently contains 1.2 lakh essential oil records with data from 92 plant taxonomic families. But, that is based on only a small set of articles and Dr. Yadav is convinced that she will be able to greatly increase this database with a search of the full scholarly corpus.

A third example was recently featured in Nature, this one in the area of materials science. The discovery of new materials is a mixture of craft and science. It is often a trial-and-error process, and is a very inefficient, almost artisanal process.

Using 3.3 million scientific abstracts, the researchers created a 500,000 word vocabulary, then looked at co-occurrences of words—such as "iron" or "steel"—and other terms, such as chemical compositions—using unsupervised machine learning. These word vectors were then associated with various materials, which were then clustered around major categories of uses, such as superconductors, battery materials, photovoltaics, and organic compounds.

This example shows the potential of data mining, but what if they had more than just abstracts to work with? Would the results be better?

A study in PLOS Computational Biology did text mining of protein-protein, disease-gene, and protein subcellular associations to examine that question.

This study compared the results from performing extraction on 15 million abstracts with the results from the same procedure on 15 million full text articles. As one would intuit, the results were far superior with full text. The reason is simple. Abstracts are highly summarized and the full text has much more detail.

Text and data mining is not just for the hard sciences. Legal informatics has used text and data mining to examine similarity in court opinions to see, for example, how U.S. District Court and Court of Appeals decisions influenced the U.S. Supreme Court.

Text and data mining is a key component of modern search engines, it is used for machine-assisted translation, it was even used recently to to determine what makes people happy!

The Economist reported on this recent study, which examined over 8 million books and millions of newspaper articles for terms with a psychological valence of happiness. Researchers found that as wealth increased people became happier, but that was incidental. Significantly more important for happiness was the health of the population and the absence of war.

Text and data mining can also be used for plagiarism detection. You may be familiar with Dr. Elisabeth Bik, the "lab fairy," who has been highlighting unattributed or manipulated reuse of imagery in articles. Image recognition on the full text of articles could substantially assist in this enterprise, as could techniques to mathematically model the words of all journal articles in order to do plagiarism detection on texts.

This clearly is an important frontier of science, but it is an unexplored frontier.

We have set out to change that.

For the past 18 months, I have been collaborating with my colleague Dr. Andrew Lynn of the School of Computational and Integrative Sciences at Jawaharlal Nehru University. Many of you are familiar with Professor Lynn's distinguished career in applying informatics to the biological sciences.

We have built a system we call the JNU Data Depot. A replica of this system is now spinning at IIT Delhi under the direction of Dr. Sanjiva Prasad, a distinguished computer scientist.

The JNU Data Depot consists of 2 large systems, each with 24 disk drives, and a cluster of smaller towers. The computers are cut off from the broader Internet, what we call an air-gapped system. We are spinning a bit over 500 terabytes of data.

On that system are the texts of over 125 million journal articles culled from a variety of sources. There is overlap in many of the sources, but we believe we have approximately 75 million unique articles.

The text has been extracted from underlying PDF files using common utilities such as pdftotext and grobid, which builds on pdftotext to add XML-based structure to the extract. Images and other components are also extracted for analysis. A number of corollary data sets are also on the system, such as the Crossref database of citations and a number of important biological databases such as the Elixir data sets.

The system contains copyrighted data, so it is carefully secured from the rest of the Internet. You need to bring your computer to campus, apply to Dr. Lynn for permission, and must agree to the JNU Data Depot terms of use. Our terms of use are a direct copy from the Hathi Trust text and data mining facility in the U.S. and strictly prohibit any redistribution of the underlying texts.

The Hathi Trust system contains all the books scanned by Google Books. Their terms of use—and ours as well—establish that the system may only be used for non-consumptive text and data mining. You are not reading or even looking at the full text of the PDF files, you are using computers to mine the text for facts.

We are limiting the system to non-commercial uses only, and have placed a number of other limitations on the system. The system was described recently in an article in Nature magazine.

India is on the frontier of this revolution, one of the first places in the world where researchers may do this type of research on a collection that approaches the full scientific corpus. We are in the early days of the system, and are still bringing it up to speed, but the system exists and is in use by a number of researchers.

☸

You may rightly ask, is this legal?

We approached the JNU Data Depot in an exceedingly deliberate and careful manner. In particular, we discussed the issues extensively with a number of legal scholars, and have put on the record a legal analysis by Professor Arul Georgia Scaria, a leading intellectual property expert at National Law University, Delhi, and by Dr. Zakir Thomas, a senior member of the civil service who was formerly the Registrar of Copyright for the Government of India, who contributed his analysis in his personal capacity.

We also consulted with a number of other legal scholars, including Professor Feroz Ali at IIT Madras, Professor N.S. Gopalakrishnan of the Cochi Institute of Science and Technology—widely considered the dean of the intellectual property community in India—and Professor Lawrence Liang of the Ambedkar University School of Law.

I also discussed the matter extensively with Professor Shamnad Basheer, who I am sad to say recently passed away at an all-too-young age, after a brilliant career of public work and public advocacy. He is sorely missed.

I am not a lawyer, but I am familiar with many of the issues because of my work making knowledge available in the U.S., Europe, and India. I am always very careful before I make a database available, and believe it is crucial that one have a strong legal grounding for any action that is taken. I strongly believe that is the case here, and the legal experts agree.

In the U.S., the question of text and data mining was extensively litigated when Google started scanning all the books of the world, and did so without first asking for permission. Google was not distributing the books, they scanned them for the purpose of text and data mining—for showing snippets to users, for driving search engine results.

Hathi Trust is the consortium of universities that provided Google with the books to scan, and received a digital copy back in return. They made the books available to their members, which are the major universities of the United States, to show their users previews and snippets, and in many cases the full text.

Hathi Trust went even further, and created a text and data mining facility where researchers may—using the same model we are using at the JNU Data Depot—mine the full text of all the books that Google scanned.

The courts have ruled repeatedly—and definitively—that the uses of both Google and of Hathi Trust are legal.

But, is this legal in India you may ask? And, are journal articles the same as books? The answer to both questions is yes.

India is a common law country. There are statutes, but ultimately any new use must be judged by the courts if somebody were to object. But we believe we stand on very solid ground and we do not expect to be challenged because of the careful and deliberate way we have gone about this, the wide-spread participation and support for this endeavor from major universities and government research labs throughout India, and the obvious and compelling need for this facility.

A basic concept in copyright law is that there is no copyright in facts and ideas, it only protects the expression of those ideas as fixed into a tangible medium. One cannot use end user license agreements to prohibit the use of facts and ideas to override the underlying purposes and rights under copyright law and the Constitution. That is morally wrong and legally wrong.

It is also important to understand that copyright law is about the rights of users as much as about the rights of creators. That was underscored in the opinion of the Hon'ble Justice Endlaw in the famous Delhi University

Copyshop case, where he said that the rights of copyright users are not be read narrowly or strictly.

Under Indian law, there are a number of exceptions to copyright. What that means is that even if a work is under copyright, certain uses are allowed. The Delhi University Copyshop case was about one of those exceptions. A professor may assign a course pack of materials that are under copyright because one of the exceptions is when materials are furnished by a teacher to a student in the course of instruction.

Another exception to copyright is for private or personal uses, specifically including research uses. That is what is happening in the JNU Data Depot.

The exceptions to copyright are part of an analysis known as fair dealing. What is "fair" is a matter of degree, it depends on the circumstances and a number of factors, and it is a matter for the courts if a controversy should arise. We hope and believe our work will not reach the courts because it is clearly fair dealing, it is clearly vital to the future of scientific inquiry, and it is in line with the goals and objectives that have been laid out by the Government of India to continue to keep India at the leading edge of the frontiers of science.

In the JNU Data Depot, only small snippets are being extracted. The purpose of our dealing is strictly non-commercial. It is being done as non-consumptive use.

Articles are not being distributed, they are not being made available to others, our use is not in any way harming the market for these works.

One of the most important factors in evaluating fair dealing is to examine the alternatives that are available. In the case of text and data mining, there is no possibility of licensing the entire scientific corpus. Yet, this corpus is what is known as an essential utility—one cannot conduct science properly without using this corpus, just as one could not operate a hospital without electricity. Even in the physical world, if a property may not be accessed without trespassing another's property, the law permits easement rights.

Some might say that keeping a copy of these materials without the permission of publishers is an infringement. But, the whole purpose of copyright exceptions is that one may keep copyrighted material if those materials are within the scope of the exceptions. The JNU Data Depot is clearly within the scope.

We strongly believe—and eminent legal experts agree with us—that this is legal.

☸

Before I conclude, I would be remiss if I did not discuss who let the elephant into this room, into our laboratories, our ivory towers. It was us. We are responsible. We are the ones that left the door open.

Since the beginnings of journals, back to the Philosophical Transactions of the Royal Society and before, the dissemination of scholarly results has been a collective endeavor.

When we present a lecture at a conference, this is not an experience for which we demand a fee. When we peer review a colleague's paper, we do this as a service to our profession. When we submit a paper to a journal, we do so for fame and glory, or to promote the increase and diffusion of knowledge, not for bags of gold.

Until recently, the editing of journals was always a matter of prestige and learning. Even today, it is almost always done for the advancement of science, not for the pursuit of pecuniary riches.

For the last 50 years, journal prices have skyrocketed to absurd heights even as costs have plummeted. The academic publishing "industry" has engaged in predatory practices, the oligarchy has built the walls ever-higher.

This is our fault.

When we complain, the elephants tell us there is no free lunch. They say the process is by its nature exceedingly expensive and complicated, they say publishing is difficult, they say we must not sacrifice quality by permitting amateurs into their walled garden.

This is nonsense.

A.J. Liebling once said “freedom of the press belongs to those who own one” and in today’s world of the Internet and computers, we all own a printing press, we all have ready and easy access to the means of distribution.

Publishers are now pretending to embrace openness, but they are doing so with smokescreens. Article publication charges are ridiculously high, lock-down periods are unnecessary and in many cases illegal, exclusive publishing platforms with abusive terms of use are a travesty.

This is not open. Publishers are attempting to fool us with their jadoowallah routines.

But, I am pleased to report that the elephants are on the run.

Plan S is a great start. The funders of science have said this must all change. They have said enough is enough. Artificial metrics are a lazy alternative to judging individual articles on their merits. Open publishing must be a requirement, not an alternative. Plan S makes it clear across the globe that change must come.

I was delighted to see the strong support of Plan S from Dr. VijayRaghavan and the Government of India. He joins leaders from around the world in supporting your efforts as scientists and scholars.

But, we must go further.

We must not depend on scholarly merchant houses for the dissemination of knowledge. As scholars, we must take back control of science. If a high-quality open journal does not exist, make one. If you publish your results, you must do so openly. There is no excuse to do otherwise.

I have great hope for the future of open publishing. The wind is at our backs. People like Dr. VijayRaghavan are helping row this boat. It will happen.

But, we must still stand on the shoulders of giants if we wish to reach new heights. This is why text and data mining on the existing corpus is so important. This is why we have created the JNU Data Depot.

☸

You are no doubt familiar with Mahatma Gandhi's seven social sins. Science without humanity is of course one of those sins. One of the great things about science in India, for the most part at least—and particularly so with those of you gathered here today—is that science is practiced with humanity. So many of you are working to solve pressing problems that afflict India and the world. It is inspiring.

But, another sin is commerce without morality, and I put it to you that that prohibiting text and data mining—telling a scientist who believes she can better find a cure for a disease or understand the roots of poverty—that is the very embodiment of that sin. Gandhi Ji would have been aghast at this unholy proposition.

Today, knowledge has been colonized. You are being told you cannot make knowledge without paying a tax and securing a license. How is this different than being told you may not make salt without paying a tax and securing a permit?

Knowledge and salt are both vital to the functioning of society, both are essential to human life. Taxing knowledge and prohibiting its consumption, taxing salt and prohibiting its consumption, these are both examples of commerce without morality.

You are no doubt also familiar with Gandhi Ji's use of the concept of swadeshi, one he of course adopted from the works of many others throughout India. Most people think of this in terms of Bapu and his spinning wheel, making kadhi as a way of fighting the colonization of India.

This was bread labor, the idea from the Sermon on the Mount that you should "earn thy bread by the sweat of thy brow." Bread labor and swadeshi, public work and satyagraha, these are the tools the Mahatma knew would lead to swaraj, to the liberation of India.

What you may not know was that bread labor for Gandhi Ji was originally not the spinning wheel, it was the printing press. At the Phoenix Ashram, their bread labor was that everybody must typeset every day. It was the increase and diffusion of knowledge that would educate the people.

Before satyagraha must come awareness of their condition among the people, and awareness comes from knowledge. Only with knowledge could the shackles of the Raj be removed.

Today, it is knowledge that has been colonized and I believe we cannot ignore this elephant in the room. We cannot sit back and allow others to say that it is wrong to pursue knowledge, that it is wrong to try and alleviate poverty, that it is wrong to make technologies that are appropriate to our modern lives and the problems our society face, that will help enlighten humanity.

Gyan swaraj is a great challenge of our times. It is a true frontier of science. You are all public workers, you have devoted your careers to the pursuit of knowledge, to the betterment of our society, to educating our youth.

If there is a way to carry out that mission better, we must not shirk from that task. If we all stand up and say text and data mining is an important frontier, then it will be so. But, it will only be so if we do this together.

Jain Hind. Jai Gyan.

Thank you very much.

About the JNU Data Depot
https://law.resource.org/pub/in/jnu/

The Hind Swaraj Collection
https://archive.org/details/HindSwaraj

The Servants of Knowledge Collection
https://archive.org/details/ServantsOfKnowledge

The Public Library of India Collection
https://archive.org/details/digitallibraryindia

The Indian Culture Collection
https://archive.org/details/IndiaCulture

The Indian History Collection
https://archive.org/details/IndiaHistory

Code Swaraj (Available in 10 Languages)
https://archive.org/details/CodeSwaraj

The Jai Gyan Top-Level Collection
https://archive.org/details/JaiGyan

Videos of Talks In and About India
https://www.youtube.com/user/LawResourceOrg/

Carl on Twitter
@carlmalamud

TALES OF A TELUGU PARIAH TRIBE
WHILE SEWING SANDALS
CLOUGH
FORTY-TWO YEARS AMONGST THE INDIANS AND ESKIMO
BEATRICE BATTY
VOYAGE TO THE ISLE OF FRANCE
THE ISLE OF BOURBON AND THE CAPE OF GOOD HOPE
1768-1771
PIERRE
THE STORY OF LANKA
L. E. BLAZE
CAWNPORE
1857
THE STATION
THE OUTBREAK
THE SIEGE
THE TREACHERY
THE MASSACRE
TREVELYAN
ALBUQUERQUE
THE EARLY PORTUGUESE SETTLEMENT IN INDIA
H. MORSE STEPHENS
THE WILD ELEPHANT & THE METHOD OF CAPTURING AND TAMING IT IN CEYLON
J. EMERSON TENNENT
THE LIFE AND LETTERS OF THE RIGHT HONOURABLE FRIEDRICH MAX MÜLLER
IN TWO VOLUMES
VOL. II
EDITED BY HIS WIFE
LIFE AND WORK IN BENARES AND KUMAON
1839 TO 1877
BENARES
KUMAON
ALMORA
RANEE KHET
JAMES KENNEDY
TARA
A Mahratta Tale
1657
MEADOWS TAYLOR
THE MYSTERIES OF NATURE
THE SCIENCE OF BREATH
PATRA
SOME SAYINGS OF THE BUDDHA
F. L. WOODWARD
A JOURNEY TO KATMANDU
OLIPHANT
THE ANDHRAS
AN HISTORICAL SKETCH OF GOA
COTTINEAU
A NOBLE QUEEN
A ROMANCE OF INDIAN HISTORY
MEADOWS TAYLOR
BUDDHIST PHILOSOPHY IN INDIA AND CEYLON
A. BERRIEDALE KEITH
THE BUDDHIST CONCEPTION OF SPIRITS
BIMALA CHURN LAW
PRE-MUSSALMAN INDIA
M. S. NATESON
DRAMATIC HISTORY OF INDIA
FLORA ANNIE STEEL
BUDDHISM
ITS HISTORY AND LITERATURE
RHYS DAVIDS
THE LEGENDS AND THEORIES OF THE BUDDHISTS
HISTORY AND SCIENCE
R. SPENCE HARDY
LEGENDS OF INDIAN BUDDHISM
WINIFRED STEPHENS
PRACTICAL BAZAAR MEDICINES
G. T. BIRDWOOD
AGRA AND THE TAJ
E. B. HAVELL
BUDDHIST INDIA
T. W. RHYS DAVIDS
C. V. VAIDYA
EPIC INDIA
HAND BOOK OF COLLOQUIAL TIBETAN
PANJABI MANUAL AND GRAMMAR
A GUIDE TO THE COLLOQUIAL PANJABI
BAILEY CUMMINGS
THE GARDEN OF INDIA
CHAPTERS ON OUDH HISTORY AND AFFAIRS
H. C. IRWIN
100 MASTERPIECES
MOHAMMEDAN AND ORIENTAL
CULAVAMSA
BEING THE MORE RECENT PART OF THE MAHAVAMSA
GEIGER
SOME CONTRIBUTIONS OF SOUTH INDIA TO INDIAN CULTURE
TIBET THE MYSTERIOUS
THE POSITION OF WOMEN IN THE VEDIC RITUAL
HISTORY OF URDU LITERATURE
THE TRIBES AND CASTES OF THE NORTH WESTERN PROVINCES AND OUDH
THE JOURNAL OF JOHN JOURDAIN
1608-1617
FOSTER

No Rights Reserved

ISBN 978-1-892628-17-6
90000
9 781892 628176